

Principal Component Analysis of Price Variations in Stock Price Time Series

B. G. SHARMA

Department of physics and Computer Science,
Govt. Nagarjuna P.G. College of Science Raipur-492010, C.G., INDIA.

ABSTRACT

In this paper the time series of stock price variation is analyzed through Principal Component Analysis to study the major trends present in a time series data. It is shown that the major eigen values are valuable tools to predict the different trends of the data set.

Keywords: PCA, Time series.

1. INTRODUCTION

Principal component analysis (PCA) is one of the important methods of modern data analysis¹⁻⁴. Frye⁵ and Jamshidian and Zhu⁶ explain in detail how trading firms may use PCA as a basis for their risk management process.

The central idea of PCA is to reduce the dimensionality of a data set consisting of large number interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal component's which are uncorrelated and which are ordered so that the first few retain most of the variation present in all of the original variables. It is a simple, non-parametric method of extracting relevant information from raw data sets. It

provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that often underlie it. It reduces data dimensionality by performing a covariance analysis between factors. As such, it is suitable for data sets in multiple dimensions. PCA is appropriate when measures on a number of observed variables are obtained and it is desired to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables. The principal components may then be used as predictor or criterion variables in subsequent analyses.

2 THEORY

PCA involves the procedure that transforms a number of (possibly) correlated

variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Basically there are two objectives of PCA: first is to discover or to reduce the dimensionality of the data set and the second is to identify new meaningful underlying variables. Traditionally, principal component analysis is performed on the symmetric Covariance matrix or on the symmetric Correlation matrix. The data is first standardized if the variances of variables differ much, or if the units of measurement of the variables differ.

2.1 Determination of the principal components

Principal components are obtained by projecting the multivariate data vectors on the space spanned by the eigenvectors. The mathematical technique used in PCA is called eigen analysis: solve for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of the eigenvalues equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix.

3. DATA ANALYSIS

The data set consists of NIFTY values and inflation index. The original data are plotted in figure 1, figure 2 shows the mean subtracted data and final data obtained by PCA are plotted in figure 3.

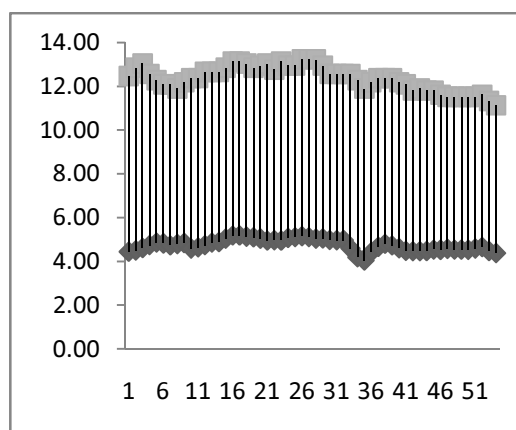


Fig.1 Plot of Original data

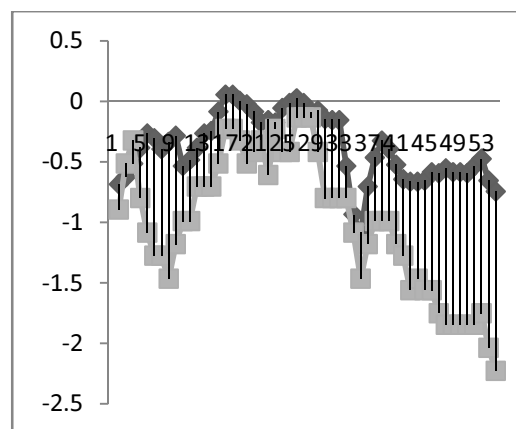


Fig.2 Plot of mean subtracted data

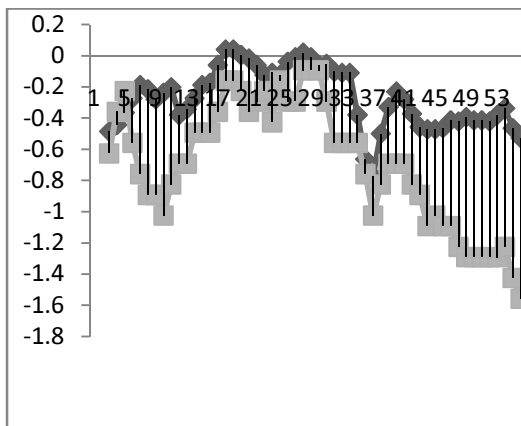


Fig.3 Plot of Final Data

4. RESULT AND DISCUSSION

PCA was conducted in a sequence of steps, with somewhat subjective decisions being made at many of these steps. In this analysis the first component can be expected to account for a fairly large amount of the total variance. Each succeeding component will account for progressively smaller amounts of variance. Although a large number of components may be extracted in this way, only the first few components will be important enough to be retained for interpretation. An eigenvalue represents the amount of variance that is accounted for by a given component. The rationale for this criterion is straightforward. Each observed variable contributes one unit of variance to the total variance in the data set. Any component that displays an eigenvalue greater than 1 is accounting for a greater amount of variance than had been contributed by one variable. Such a component is therefore accounting for a meaningful amount of variance, and is

worthy of being retained. On the other hand, a component with an eigenvalue less than 1 is accounting for less variance than had been contributed by one variable. The purpose of the analysis is to reduce a number of observed variables into a relatively smaller number of components; this cannot be effectively achieved if retains to components that account for less variance than had been contributed by individual variables. For this reason, components with eigenvalues less than 1 are viewed as trivial, and are not retained. After calculating eigen values and on arranging them in order of their significance, leaving least significant values the final data is calculated. This final data is plotted in fig.3. The data set obtained through PCA is compared with original time series. It was found that the method effectively reduces the dimensionalities of the data with a very small loss of information contained in the original time series. In this way PCA proved itself as an useful tool whenever there is large number of data's and it is desired to extract most significant data's from those data points.

5. CONCLUSION

Principal component analysis is best performed on stochastic data whose standard deviations are reflective of their relative significance for an application. This is because principal component analysis depends upon both the correlations between random variables and the standard deviations of those random variables. If it is desired to change standard deviations of a set of random variables but leave their correlations the same, this would change their principal components. PCA uses standard deviation as a metric of

significance. If one variable has a standard deviation that far exceeds the rest, that variable will dominate the first eigenvector. Unfortunately, there may be no correspondence between a variable's standard deviation and its significance. Standard deviations depend upon the units in which a variable is measured.

If principal components are used only to orthogonalize a random vector, this will not be a problem. No information is lost. It will be a problem if principal components are discarded to form an approximation. In this case, information is lost. Before it is discarded that principal components that appear "insignificant," it should be made sure that they truly are insignificant.

There are various solutions to this problem. It might insist that all variables be measured in the same units, but this is not always feasible. Also, identical units do not necessarily correspond to identical significance. Alternatively, principal component analysis is applied to normalized random variables obtained by dividing each random variable by its standard deviation. With this approach, one can effectively

apply principal component analysis to the random variables' correlation matrix. This represents a different weighting from that obtained by measuring all random variables in identical units, but not necessarily a better one.

6. REFERENCES

1. J.D. Hamilton. Time Series Analysis, Princeton, New Jersey, (1994).
2. Sprott J.C. Chaos and time series analysis, pp 260, Oxford University Press (2008).
3. Kantz, H. and Schreiber T. Nonlinear time series analysis, Cambridge University Press (1997).
4. R.A Johnson, D.W Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, New Jersey, (1988).
5. Frye, Jon. "Principals of Risk: Finding Value-at-Risk through Factor-based Interest Rate Scenarios, manuscript, Chicago: Nationsbanc-CRT (June) 39-69 (1996).
6. Jamshidian, Farshid, and Yu Zhu. "Scenario Simulation Model: Theory and Methodology." (1996).